DOCUMENT RESUME

ED 099 431 95 TM 004 309

AUTHOR Flaugher, Ronald L.
TITLE Bias in Testing: A Review and Discussion. TM Report No. 36.
INSTITUTION ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO ETS-TM-36
PUB DATE Dec 74
CONTRACT OEC-0-70-3797-519
NOTE 10p.

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Competitive Selection; *Minority Groups; Personnel Selection; Predictive Validity; Racial Discrimination; Sex Discrimination; *Test Bias; *Testing

ABSTRACT

Recent empirical evidence concerning sex and racial bias in testing is discussed in terms of three primary sources of bias: (1) content of the test itself, (2) atmosphere in which the test is administered, and (3) the use to which the test results are put. Test content that is demonstrably more difficult for one group than another should be (1) eliminated in any setting in which equal difficulty is assumed or (2) perhaps more important, the biased content should be examined closely for possible causes of the difference, leading to modification of educational practices for the low-scoring groups. Special care should be taken routinely to see that minority groups are made to feel comfortable and are not intimidated by their surroundings. Pertaining to fairness in test use, methodological developments undermining the traditional statistical model of fairness previously accepted without question are described in some detail. The "new measures" approach to test bias is seen as essentially an abandonment of, or a reduced emphasis on, the traditional measures of status of aptitude and achievement. (Author/RC)

ED 099431

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION
EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540

TM REPORT 36 DECEMBER 1974

BIAS IN TESTING: A REVIEW AND DISCUSSION

Ronald L. Flaugher

A little over four years ago, this author reviewed the literature on test bias under the title *Testing Practices, Minority Groups, and Higher Education* (1970). The present paper attempts to update the topics discussed, reviewing the research progress made, and in certain areas revising the outlook presented there. Since many of the same problems are common to employment testing, both that and educational testing are included within the following discussions.

To summarize the 1970 paper very briefly, it grouped the possible sources of test bias—or unfairness—into three categories and discussed the research findings on each.

First, and most commonly perceived as the source of bias, is the content of the test itself. Very reasonably, if a test is biased, it must be because of what is in that test. Within this category are the questions of predictive validity for minority groups. But a second category is that of the *atmosphere* in which the test is administered, the environment being an important determinant of how the student or applicant actually performs on the test regardless of content, even including whether or not the person comes forward to take the examination at all.

The third category of unfairness in testing has to do with the use to which the test results are put. In the 1970 paper I stated:

> Test use, however, is seldom regarded as a subject for at least the ordinary kinds of research effort. Unfairness from any source, however, can be the weak link in an otherwise strong chain and misguided use of test results can be a very serious defect in a testing program (p. 7).

It will be seen later in this paper that one shift of emphasis occurring is the increasing realization that test *use* may be the one most important source of unfairness, deserving a great deal more attention than the others, certainly more than it has received in the past.

The Issue of Differential Validity

In 1970, the conclusion was that perhaps some of the criticism of test content was inappropriate, based as it was on the belief that the tests *do not predict as accurately* for minorities as for the majority group. The evidence available at that time was considerable, but the strength of the conviction that differential predictive validity existed was even more considerable.

In the intervening years, the empirical evidence has continued to accumulate and continued to point in the same direction. Davis and Temp (1971) collected data on the relationship between freshman grades and SAT scores for groups of black and white students at the same colleges. They found wide variation in validities across the institutions, and a tendency for validities to be higher for white groups than black groups, although the difference was not large. They found, similar to findings reported in the earlier paper (Flaugher, 1970, p. 13), that there was a tendency for black students to be predicted to do better than they actually do when a common prediction equation is used for both black and white students. The authors reemphasize the need to keep verifying the validities on a local level for both groups. Pfeifer and Sedlacek (1971) and Kallingal (1971) also found overprediction for black students in two other settings, but stressed the need for separate prediction equations.

Schmidt et al. (1973) reviewed the results of a number of different studies in industrial settings and concluded that

> . . . for both subjective and objective criterion measures, observed frequencies of both kinds of single-group validity (significant for whites but not for blacks and vice versa) were not significantly different from those predicted by the null differences model. These findings cast serious doubt on the existence of single group and differential validity as substantive phenomena (p. 5).

Additional results are available from an extensive study by Campbell and associates (1973), which concludes that

> . . . it appears that differential validity, if not entirely a statistical artifact where it does appear, is at best an isolated phenomenon (p. 425).

Meanwhile certain legal activities may be prolonging the vitality of the conviction that differential validity is an issue. The federal government is attempting to produce

TM 004 309

This publication was prepared pursuant to a contract with the National Institute of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, represent official National Institute of Education position or policy.

new guidelines for employment practices in American industry, and one reason for the prolonged delay in producing them has been the issue of differential validity, the existence of which is still considered to be a matter of debate (Singer, 1974). Substantive research findings are once again being challenged by what "everyone knows," regardless of the data.

Regardless of the evidence to indicate that the tests can do as good a job of prediction for minorities as for the majority group, the specific content of the tests has come under continuing scrutiny. Angoff and Ford (1971) conducted item analyses for both black and white high school students, and found that the item-by-group interactions could be reduced considerably by relatively crude matching techniques, suggesting that the interactions might disappear altogether with more careful matching on performance level. Further, they found significant item-by-*city* interactions for black students, casting doubt on the existence of a single body of content that, if included, would be uniformly advantageous to black students.

Certainly particular test content that is demonstrably more difficult for one group than another should either be 1) eliminated in any setting in which equal difficulty is assumed or 2), perhaps more important, the biased content should be examined closely for possible causes of the difference, leading to suggestions about modifying educational practices for the low-scoring groups. Thus, in an extensive study by Breland et al. (1974), the performances of 10 sociocultural groups over six cognitive tests were analyzed for instability of difficulty level across groups. Quoting Breland et al.:

> The greatest instabilities were noted among the vocabulary items. These vocabulary instabilities appeared to be attributable to linguistic differences, primarily those existing between Spanish-speaking groups and other groups. It was also observed that reading test items having material relevant to black culture were relatively easier for blacks than were other items in the test battery. A perhaps significant finding occurred in the analysis of mathematics items. Mathematical knowledge obtainable from everyday life situations, such as how to count money, were relatively less difficult for minority groups [than other classes of items]. In contrast, very simple mathematical problems, such as determining the value of square roots of whole numbers less than ten, seem extraordinarily difficult for minority groups. Since such knowledge, though easily obtained, is usually only obtained in a school setting, what is suggested is that most minority groups in the United States receive seriously deficient schooling in mathematics (p. iii).

In another, more unique approach to the question of biased test content, Green (1972) suggests separate application of the same test construction techniques for each ethnic group, using a single common pool from which items are to be selected for inclusion in a test. To the extent that the selected items overlap for the various groups, the test may be said to be unbiased. Green's empirical findings would indicate that indeed the overlap of selected items is frequently less than perfect, and his technique can be a source of information about the impact of test content interacting with the ethnic group identity of the test taker.

In general, however, as with other related studies of test content, the nature and size of the changes that would be made in test content on the basis of these data are such that only small differences would be realized in the test scores received by individual test takers. This conclusion is eventually drawn by most lines of investigation in pursuit of the topic. The findings are important for greater understanding of both testing and the nature of educational opportunities for minorities, but mere elimination of some small subset of test items does not appear to be the answer to the much more sizeable problem being referred to as test bias.

The Atmosphere of Testing

In the second general category of unfairness discussed in the 1970 paper, namely atmosphere, very little substantive research has been forthcoming that would change the conclusions made in that paper. Additional studies of such peripheral characteristics of the test as time limits have been conducted, such as that by Evans and Reilly (1972) on the Law School Admission Test. Their conclusions are typical:

> 1) The test is somewhat more speeded for [predominantly black groups] than for regular candidates, 2) reducing the amount of speededness produces higher scores for both [groups], and 3) reducing speededness is *not* significantly more beneficial to [the predominantly black groups] (p. 123).

Another aspect of the atmosphere of testing is the amount of sophistication or experience needed to overcome the idiosyncratic characteristics of the testing situation, including such things as the type of test item and the answer sheet format. To give the most favorable representation of their abilities, students must overcome the *medium* and concentrate on the *message*, the test content itself. Thus, some people are characterized as being able to take tests and others as not, independent of their competence with the subject matter. The middle-class student is seen by minorities as being able to take tests because of greater experience with the "tricks" required to perform well.

There is some evidence that test wiseness amounts to an ability to take advantage of the violations of good item-construction principles (McMorris et al., 1972), suggesting that more carefully constructed measures are less susceptible to these effects, although positive results have been achieved with elementary-age students on a

standardized reading test (Callenbach, 1973). A very careful and intensive study on high school students by Evans and Pike (1973) did achieve positive results for three types of mathematics items, but the 21 hours of instruction over the seven weeks of the experiment were of such a nature and intensity that they might be regarded as a legitimate mathematics curriculum rather than a content-free coaching session.

Epps (1974) has continued the work reported at length in an earlier paper by Katz (1970) and, in general, continued to find that atmosphere variables, such as the race of the examiner and the perceived use to which the test results were to be put, usually have a detectable effect on test performance and motivation of minorities and, perhaps, the majority group members as well. So many atmosphere variables exist and remain unstudied, however, that little progress has been made in untangling the multiple interactions of such things as personalities, achievement motivation, perceived comparison groups, perceived likelihood of success, and perceived status of the examiner.

In spite of incomplete research evidence, the general conclusion on the question of atmosphere effects is that special care should be taken routinely to see that minority groups are made to feel comfortable and are not intimidated by the surroundings. Such things as unusual distances of travel to the testing center, recruitment publications that discourage minorities, and insensitive treatment at the center are being seen as potentially important influences on the test performance of minorities and, therefore, to be attended to regardless of the lack of extensive firm evidence that they make a difference. Similarly, although no evidence existed at the time to indicate that it makes for better performance, test makers began including such things as reading passages by and about black people in the hope that some beneficial effect would be realized. Breland et al. (1974) later found some justification for this practice, as described earlier, though the effective difference on total test scores seems to be too small to be detected except in very large samples.

Sex Bias in Testing

Women are not the usual sort of minority group and do not have the usual sort of difficulties with testing. It is frequently the case, for example, that instead of earning lower test scores, the women in a particular sample may score better than their male counterparts, but frequently they are nevertheless restricted to filling a certain percentage of the available openings or are eliminated or discouraged on some other grounds. Sometimes when the openings for women are very desirable and scarce, this leads to large discrepancies between the average aptitude test scores for men and women, providing hard evidence of the existence of discriminative practices. These practices are being challenged and gradually being abandoned, along with the customary preference shown to men in promotions to higher levels in business and industry.

The tests, however, if not primary instruments of discrimination against women, are a source of great irritation and perhaps even real unfairness, because of the sexism in the image of women that is projected by the language. Although the nature of the language is not the responsibility of the test makers, publishers of all kinds, including those of tests, are under increasing pressure to eliminate the practice of the preferred masculine pronoun and the dominance of masculine referents from their products ("*man*kind," "fire*man*," "the average American drinks *his* coffee black"). Controversy continues over whether or not satisfaction of this demand does such violence to the usual linguistic habits that it distracts attention from the task of the test itself. McGraw-Hill (undated) has provided guidelines for its publications that solve much if not all of the anticipated awkwardness. To the extent that such changes can be made, testing materials as well as teaching materials, reference works, and nonfiction works in general can be legitimately asked to eliminate the role language has played in reinforcing the existing inequality between the sexes.

Test Use: Fairness in Statistical Selection Models

It is on the topic of *test use* that the most dramatic developments have taken place since this writer's earlier review. Additional research efforts were called for at that time, necessarily not of the usual sort, to study the fairness of the utilization of test information. It was not anticipated at that time that methodological developments were to be next—ones that would undermine the traditional statistical model of fairness previously accepted virtually without question. In the next few pages, these developments are described in some detail, inasmuch as they are quite possibly of more ultimate import than are the previously described studies of differential validity and unfairness in testing atmosphere.

Earlier in this paper, the evidence for underprediction of minority group members was found to be virtually absent in that the existing tests were found to predict about as accurately for minorities as they do for the majority group. This approach to the determination of fairness is based on a statistical model that has come to be known as the traditional, or Cleary, model, after the researcher first employing it in a study of test bias (1968). In 1971, however, both Thorndike and Darlington separately showed that the traditional definition has difficulties. Thorndike showed that regardless of the equivalence of the relationship between test and pre-

dicted criterion for the two groups, such a statistical model *is unfair to the lower scoring group of applicants, in the sense that the proportion of that group that qualifies on the test will turn out to be smaller than the proportion who would be qualified on the job.*

Certainly that definition, stated as it is in terms of proportions selected versus proportions who would succeed, seems to be reasonable and desirable. But still another definition was soon advocated by Cole (1973), who preferred to look at the situation in this manner: *Given one member of the majority group and one member of the minority group, both of whom would succeed if selected, the procedure is unfair unless they have the same probability of being selected.*

That definition seems as appealing as either Thorndike's or Cleary's, but all three are in conflict with each other and cannot be advocated or practiced simultaneously. Darlington's (1971) conceptualization of the problem is seen to be the most accurate description of all the models simultaneously, because he demonstrated that all three definitions could be encompassed within a model that retained a single, but variable-weight, correction factor. The particular size of that factor was to be determined subjectively, that is, on the basis of other factors that lie completely outside the statistical model itself. This variable factor was to be added to the criterion scores (job performance, measures of college success) of the lower scoring group, and the size of the factor was to be determined by the particular set of chosen values the selector or the selecting institution wished to invoke. No longer was it to be possible to claim that the objective statistics used in selection were the court of last appeal, the ultimate determination of just what was fair; rather, the objective procedures were seen to be a strictly mechanical implementation of the definition of fairness that had been *chosen* by the selector.

Must We Have a "Quota System" To Be Unbiased?

Up to this point, largely for the purposes of ease of presentation, this values and selection fairness discussion has proceeded as if there were a clear path from our previous naive practices to the adoption of the enlightened corrected criterion model, leading to greater equity and mutual understanding. This is not the case. The focus of the difficulty is upon the correction factor of the new formulation. When one group's scores, either on the criterion or on the selection test, are treated differently than those of the other group, this amounts to establishing a *double standard*. It is precisely the same practice in concept as establishing two different cutting scores on a selection test and the same, in effect, as deciding upon a *quota* for one or the other of the subgroups. Particularly when stated in terms of a quota system, the result is often a strong negative reaction on the part of many observers. In such cases, the one-value system, which endorses this sort of fairness in selection, takes second place to one that endorses strictly equivalent selection standards regardless of ethnic identity. All of the newly developed models of selection fairness have this perceived flaw, in that they require some correction of one group's scores to the disadvantage of the others. Only the traditional model applies no correction factor.

This dilemma is not confined to uninformed public opinion; it extends into our courts and into our declarations of public policy as well. The United States Supreme Court is currently on record as endorsing only the traditional model, implicit in their declaration that "race, religion, nationality, and sex become irrelevant" (*Griggs vs. Duke Power Company*, 401 U.S. 424), thereby ruling out differential score correction based on race. On the other hand, a more recent decision in the Supreme Court of the State of Washington has ruled that, at least in educational selection practices, racial distinctions *can* be made for *compensatory purposes* (*DeFunis vs. Odegaard*, see Ginger, 1974). Until the matter is finally ruled on by the United States Supreme Court, however, the issue is likely to be seen as legally unsettled.

This conflicting attitude toward compensation causes a great deal of disagreement about which particular model is appropriate. If it is acknowledged that an injustice has occurred in the past to a particular subgroup of the population such as an ethnic minority, it may or may not follow that some sort of compensation on the part of society is appropriate. If compensation does follow, a public policy like that expressed by the late Lyndon Johnson would be applicable:

> To be black in a white society is not to stand on level and equal ground. While the races may stand side by side, whites stand on history's mountain and blacks stand in history's hollow. Until we overcome unequal history, we cannot overcome unequal opportunity.
>
> Not a white American in all this land would fail to be outraged if an opposing team tried to insert a twelfth man in the football lineup to stop a black fullback on the football field. Yet off the field away from the stadium, outside the reach of the television cameras and from watching eyes of millions of their fellow men, every black American in this land, man or woman, plays out life running against the twelfth man of a history that they did not make and a fate they did not choose (*New York Times*, December 26, 1972).

On the other hand, if the posture taken is that whatever injustice may have occurred in the past is no reason for an injustice of another sort in the present, then the policy should be one like that of former President Nixon:

> In employment and in politics, we are confronted with the rise of the fixed quota system—as artificial and unfair a yardstick as has ever been used to deny opportunity to anyone (*Time*, October 9, 1972).

> The way to end discrimination against some is not to begin discrimination against others (Miami Beach, August 23, 1972; *Newsweek*, September 18, 1972).

Clearly, the question of test bias has led us into areas far more complex than are capable of resolution with ordinary sorts of testing research.

Test Use: Irrelevant Selection Standards

There is still another sense in which test use is a matter of concern, apart from the debate over the meaning of the various statistical models. If people are being hired for a job that requires very little use of vocabulary, then it is not appropriate to select those applicants to be hired solely on the basis of a test of vocabulary. If there is no reason to suppose that the possession of a high school diploma is necessary to the successful accomplishment of a particular job, then non-high-school graduates should not be excluded from consideration for that job.

These are seemingly quite acceptable statements, but in fact their violation is evidently widespread. Add to this the facts that a greater percentage of minority groups score poorly on vocabulary tests and more of them fail to graduate from high school, and the result is a situation that, intended or not, constitutes an effective means of discriminating against minority groups, awarding them fewer opportunities to prove themselves on the job, and perpetuating a lower standard of living.

A vocabulary test may be a perfectly legitimate, carefully constructed and administered test, but one intended for, say, prediction of success in college English courses rather than toward deciding who would make a good fire fighter. Being used as described, however, there is no question that the test use is *biased* and biased in particular against minority groups. If that same test were to be employed in a college, however, then the bias in that test would be absent. The bias, in other words, can exist totally in the misuse of the test rather than in some internal or peripheral characteristics of the test itself.

Boehm (1972) looked at 13 recent studies dealing with Negro-white differences in employment and training selection procedures. She found that 100 of the 160 validity coefficients were not significant for *either* group. "The use of nonvalidated methods for selection," she stated, "is apparently not uncommon" (1972, p. 37), and she pointed out that this frequently excludes a disproportionate number of Negroes for reasons unrelated to job performance.

The Call for New Measures

By way of summarizing to this point, there is ample evidence that if we mean by bias that the tests do not predict as accurately for one group as the other, then the tests are *not* biased when they are used in the appropriate settings. In addition, we have seen that the search for biased content in tests themselves has been rather unproductive and frequently leads instead to questions about the quality of previous education. In spite of these developments, the accusations continue, and the popular impression remains that the tests are biased against minorities. In addition, it is sometimes claimed that what is needed to overcome this bias are "new measures." The term "nontraditional means of assessment" is also heard, and that "means of fairly measuring the amount of knowledge retained by the student regardless of his or her individual background" need to be developed.

This general line of criticism originates in the conviction by many that minority groups possess talents and attributes that are unique to their groups and are valuable and important, and yet these attributes are completely absent from those tests developed by and for the dominant white majority (Blake 1971; Brazziel 1972; Cameron 1970). This, then, is another sense of the word *bias*—that the tests may be accurate and appropriate for some segments of the population, but they are not aimed at, and, therefore, cannot document, the unique attributes of people from minority cultures.

The primary difficulty with this approach, of course, is the fact that if the test score is to be useful, it must be related to some alternatives or actions in the society, such as providing a prediction of success or failure at a job or in additional educational pursuits. If certain talents remain undocumented within minority groups, a possible reason is that they are yet to be considered sufficiently important to society to document. In the current changing atmosphere, the priority for attending to undocumented talents may be upgraded; however, for the most part, research on this aspect has been slow and confined largely within the existing parameters of human aptitude technology, seeking to find new patterns of known aptitudes rather than striking out in search of entirely new ones.

In the earlier review, for example, this author discussed the findings of Lesser, Fifer, and Clark (1965) and Stodolsky and Lesser (1967), which documented the existence of differential patterns of ability in several minority groups. These same patterns were at high levels or low levels depending on the socioeconomic status of the children within the ethnic group, but they demonstrated their viability by remaining constant within those groups. These studies were of first graders in New York and Boston schools. Flaugher (1971) analyzed some available data from four similar ethnic groups on four tests taken by inner-city eleventh graders in Los Angeles, and the resulting patterns strengthen the belief that such patterns do characterize the groups. This can amount to stereotyping the various ethnic groups, however, and has given rise to fears of educational resegregation on the basis of strengths rather than ethnic identity, with the results being the same second-class treatment for minorities. Thus, the fears of how the test data might be

used to discriminate between the minorities and the majority group comes in direct conflict with the need for special treatment of the uniqueness of the minorities. Lesser, for one, has become frustrated by these developments:

> In the light of these objections, fewer and fewer young social scientists are undertaking research on culture and cognition, because the penalties to this research are all too obvious. We are facing a self-declared moratorium on research . . . about the connections that exist between cultural conditions and cognitive growth (1972, p. 24).

One approach, in its beginning state, is advocated by Gordon (1974) and it appears to be an important response to the general call for new measures. Gordon is proposing a fundamental shifting of the focus of testing. Current tests, he states, are all directed toward assessing the *status* of a person at a particular time, whether the test's aim is describing aptitude or achievement. Knowledge of a person's status is sometimes useful, of course, but an important aspect of that person remains undescribed, that of the *processes* the person has used to arrive at that status. Far more important to an educator interested in assisting a student, for example, is the nature of the processes that are being used for learning rather than the status of an individual at any particular time. When the question of minority education is addressed, the answers—the solutions—are in terms of appropriate diagnosis and prescriptions rather than prediction, and this is best done in terms of *process* variables rather than descriptions of status.

The new measures being proposed, then, are not simply extensions of the existing categories of aptitude and achievement; they are an entirely new class of measures directed toward aspects of human behavior largely neglected until now. The little previous work that exists includes that of Herbert Birch and his associates (Thomas, Chess, and Birch, 1968) and the earlier attempts by Else Haeussermann (1958). The Thomas, Chess, and Birch studies have described nine categories of temperamental traits that Gordon (1974) suggests are a good starting point for these studies: 1) activity level, 2) rhythmicity, 3) approach withdrawal, 4) adaptability, 5) intensity of reaction, 6) threshold of responsiveness, 7) quality of mood, 8) distractibility, 9) attention span and persistence. Further, Haeussermann developed an interview technique in which she described the style of learning and its developmental level, with the purpose of prescribing the best possible methods of instruction for a child. From these elemental beginnings, Gordon hopes to increase the understanding and use of these approaches to assessment. The "new measures" approach to the problem of what is wrong with tests, and, thus, the problem of test bias, is, then, essentially an abandonment of, or at least a reduced emphasis on, the traditional measures of status of aptitude and achievement.

Have We Really Been Talking about Bias?

The review up to this point has covered the definitions of bias that have to do with the content, atmosphere, and the ways in which a test can be used in both fair and unfair systems of selecting and the prospects for nontraditional means of assessment.

But there are some critics of testing who would deny that the real bias problem has been discussed at all. To those, not only are all the statistical studies totally beside the point, they may often amount to attempts to confuse the issues or divert attention from where the real problems lie. To them, the existence of culturally biased tests is not in dispute but is simply a known and established fact.

In a sense, the statement about the cultural bias is true, because certainly the culture *is* reflected in that culture's tests, and, in circumstances such as licensing examinations, the test is often the *only* tangible indication of the standards and requirements for that aspect of the culture. If the culture itself is perceived as biased against minorities, then certainly its examinations can be expected to be biased as well. If this is the sense in which test bias is used—that is, as part of a more general belief that the entire society is biased—then a case can be made that the problem lies not with the test but in the nature of the entire culture. Many spokesmen are likely to deny this position as legitimate, however, because, since tests serve as the gatekeepers for much of society's rewards, the tests make an appropriate first target in the battle to create a more just society.

To some minority spokesmen, the description of validation statistics showing that minority groups are predicted as well as the majority group and the essentially negative results from studies of biased content are no response at all to their criticisms of testing, no matter what the final results. Rather, no other statistical evidence is needed than the underrepresentation of minorities in the meaningful, lucrative, and prestigious positions in the society. Society is seen to be treating the minorities poorly, and the testing industry is a visible and active component of that poor treatment.

Robert L. Williams, for example, has described the following circumstances, which he claims characterize the educational fate of minority students. First, as the minority student initially enters school, he very likely comes from such an impoverished environment that he is less prepared for school than the typical white student. The early testing results reflect this fact, and the lower test scores are used to justify the tracking of such students into "special" classes, which are special in that they amount to the abandonment of effort for those students. Instead of the *extra* attention such test results might seem to call for, there is, in fact, *less* effort expended and following that, of course, less expectation of success. Mercer (1973) has provided a careful docu-

mentation of this process in the California schools. Sure enough, when the next round of testing is conducted, the special students score even more poorly than before. This, in turn, justifies further negative decisions about the usefulness of additional educational effort, culminating in a denial of access to a college education because the student is "unprepared."

The tests are initially used to condemn the students to an inferior education, are then used to document that same fact, and then finally deny further opportunity to the same students. And *that*, says spokesman Williams, is what is biased about testing (1974). No validity coefficients or elaborate item analyses of tests are really addressing this basic question at all, and so no amount of such data will ever convince the spokesmen that tests are, in fact, fair and that the criticisms are unfounded.

So testing is biased, our educational system is biased, our employment practices are biased, our entire society is biased. As part of this larger biased network, testing cannot, therefore, attempt to evade the blame by claiming that it only reflects the biased nature of the rest of society. But what about the validity studies on the predictive power of the tests? Doesn't this show that the tests are doing exactly what they are claiming to do? When low-scoring students, minority or majority, are admitted to a demanding college curriculum, they tend to fail in large numbers. What is biased about an accurate prediction?

A minority spokesman's reply to this might well be that the tests represent the first barrier, the first of many between minorities and successful participation in society. Once this first barrier is eliminated, then the next one will be dealt with. If this turns out to be the inability of the colleges to take the minority students as they are and educate them, then the next target will be that fact, and steps will be taken to deal with *that*. Tests are not the only barrier, but they are frequently the first, and hence the first that should be dealt with.

This position on the part tests play in allocating educational opportunities has necessarily undergone some reexamination in light of declining enrollments in the accredited colleges of the nation—there are currently thousands of empty seats available and colleges are welcoming anyone with virtually any test scores. Many colleges are in desperate financial condition because of declining income from tuition while costs are undergoing inflation, which means that more students who fill those seats must pay their own way rather than hope for a subsidy from the college. And since minority group members are very likely to have limited financial means, they are unable to take advantage of the opportunity. So the effect is the same—higher education is being denied the minorities, but now the reason is financial rather than biased testing.

There are still many highly selective colleges that have far more applicants than available spaces, however, so this characterization does not apply uniformly across colleges, and in such settings, test scores are still an important part of the selection process. Thus, questions of test bias might still arise, but in such settings, predictive validities constitute a more relevant response. But to describe the primary barrier to minorities as one of testing, while ignoring the very considerable financial barriers that, in fact, are on the increase, seems unrealistic and misleading.

What about a Moratorium on Testing?

Some individuals and groups feel so strongly about the damaging effects of testing that they have advocated a moratorium on testing. At the 1974 meeting of the NAACP, for example, a resolution was adopted that demanded "a moratorium on standardized testing wherever such tests have not been corrected for cultural bias. . . ."

Advocates of testing, however, worry about the consequences of *not* testing, citing the necessary return to subjective impressions gained from interviews, a procedure not likely to favor minorities, and the lack of a yardstick by which the educational establishment can be held accountable for the job it is doing (Messick and Anderson, 1970). In any case, it remains unclear just what impact these resolutions may have on testing. One moratorium-like step has been imposed by the test publishers themselves (Smith, 1974) involving the use of the National Teachers Examinations scores to determine salary levels for practicing teachers in South Carolina. Since the examinations are intended as measures of academic achievement rather than teaching performance itself, this constitutes a misuse, and the publishers have refused to report scores to the State Department of Education until the practice stops.

A reply to the advocates' fears about not testing is that of Gunnings (1971) who states his opinions on the consequences:

> If standardized tests were not used, I do not feel that there would be an increase in discrimination, but a decrease. No longer would a prospective employer have the excuse that Blacks are unqualified but would possibly hire the Black persons and let their actual performances on the job be the test. I further do not feel that the personal interview method of appraisal is the evil that it is thought to be. In a personal interview, such things as motivation, enthusiasm and desire for the job or determination to succeed can be detected. These are great determining factors as to whether or not one will be successful in his work (p. 76).

Thus, the return to subjective standards is evidently seen as an advantage to minorities. It is an ironical historical development that objective testing, once seen as the factor that permitted *meritocratic* principles to predominate over the previous *aristocratic* means of

selection and was thus the benefactor of minority groups, is now seen as a barrier that must be overcome by them (Cross, 1971, p. 2).

Concluding Remarks

At the beginning of this paper, it was stated that a shift of emphasis has occurred since the first attempt to deal with the question of test bias. By now perhaps it is clear that there has been an increasing realization on the part of those concerned with testing that a test *cannot* be biased in the abstract; it must be biased or unbiased in a particular use. Regarding an IQ test as a fixed, culturally fair measure of a person's "raw" ability rather than as a reflection of culture-bound achievement over time is an example of bias. The test itself, however, cannot be either biased or unbiased. Requiring a high school diploma for a job may be biased in its effect, on the other hand, even when the intent of the selector is only to select those who will do the best job. Certainly test content and testing atmosphere should be constantly explored for indications that any subgroups are being handicapped unfairly, but by far the most dangerous—and the most difficult—source of bias is what people do with the information, intentionally or not. Test scores cannot, at present, measure many of the most highly valued aspects of human behavior and are not likely ever to do so. Motivation and creativity, however defined, are simply too elusive for standardized measurement. Further, test scores are all subject to change, sometimes by dramatic amounts, and therefore should never be regarded as immutable. To do poorly on a test is not to be condemned forever to society's reject pile. The personal worth of an individual is not summarized in an IQ score (Flaugher, 1974), even though the public seems to want to overinterpret it this way.

Misuse of test information, then, has a widespread and significant impact on the lives of minorities, even as it is being acknowledged that such information is necessary to maintain educational accountability by an objective, publically agreed-upon standard. Test makers and interpreters alike must be held responsible for proper interpretation of the measurements, to a degree far greater than has been the case in the past.

REFERENCES

Angoff, W.H., and Ford, S.F. *Item-race interaction on a test of scholastic aptitude.* Research Bulletin 71-59. Princeton, N.J.: Educational Testing Service, 1971.

Blake, E., Jr. Test information as a reinforcer of negative attitudes toward black Americans. In *Proceedings of the 1970 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1971.

Boehm, V.R. Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. *Journal of Applied Psychology*, 1972, 56(1), 33-39.

Brazziel, W.F. School testing and minority children. Summary of address prepared for presentation at conferences on test bias sponsored by National Education Association, February 18-20, 1972, Washington, D.C.

Breland, H.M., Stocking, M., Pinchak, B.M., and Abrams, N. *The cross-cultural stability of mental test items.* Project Report 74-2. Princeton, N.J.: Educational Testing Service, 1974.

Callenbach, C. The effects of instruction and practice in content-independent test-taking techniques upon the standardized reading test scores of selected second-grade students. *Journal of Educational Measurement*, 1973, *10*(1), 25-30.

Cameron, H.K. Cultural myopia. *Measurement and Evaluation in Guidance*, 1970, *3*(1), 10-17.

Campbell, J.T., Crooks, L.A., Mahoney, M.H., and Rock, D.A. *An investigation of sources of bias in the prediction of job performance: A six-year study.* Project Report 73-37. Princeton, N.J.: Educational Testing Service, 1973.

Cleary, T.A. Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 1968, *5*, 115-24.

Cole, N.S. Bias in selection. *Journal of Educational Measurement*, 1973, *10*, 237-55.

Cross, P. *Beyond the open door: New students to higher education.* San Francisco: Jossey-Bass, 1971.

Darlington, R.B. Another look at cultural fairness. *Journal of Educational Measurement*, 1971, *8*(2), 71-82.

Davis, J.A., and Temp, G. Is the SAT biased against black students? *College Board Review*, 1971, 81, 5-9.

Epps, E.G. Situational effects in testing. In Miller, L.P. (ed.), *The testing of black students: A symposium.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1974.

Evans, F.R., and Pike, L.W. The effects of instruction for three mathematics item formats. *Journal of Educational Measurement*, 1973, *10*, 257-72.

Evans, F.R., and Reilly, R.R. A study of speededness as a source of test bias. *Journal of Educational Measurement*, 1972, *9*(2), 123-31.

Flaugher, R.L. *Testing practices, minority groups, and higher education: A review and discussion of the research.* Research Bulletin 70-41. Princeton, N.J.: Educational Testing Service, 1970.

Flaugher, R.L. *Project Access research report No. 2: Patterns of test performance by high school students of four ethnic identities.* Research Bulletin 71-25. Princeton, N.J.: Educational Testing Service, 1971.

Flaugher, R.L. Some points of confusion in discussing the testing of black students. In Miller, L.P. (ed.), *The testing of black students: A symposium.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1974.

Ginger, A.F. (ed.). *DeFunis versus Odegaard and the University of Washington.* Dobbs Ferry, N.Y.: Oceana Publications, Inc., 1974.

Gordon, E.W. Affective response tendencies and self-understanding. In *Proceedings of the 1973 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1974.

Green, D.R. *Racial and ethnic bias in test construction.* Monterey, Calif.: Calif. Test Bureau/McGraw-Hill, 1972.

Griggs vs. Duke Power Company, 401 U.S. 424.

Gwaltney, T.S. Response to critics of Robert I. Williams. *Counseling Psychologist*, 1971, *2*(3), 73-77.

Haeussermann, E. *Developmental potential of preschool children.* New York: Grune and Stratton, Inc., 1958.

Johnson, L.B. Running against the twelfth man of history. *New York Times*, December 26, 1972, 33.

Kallingal, A. The prediction of grades for black and white students at Michigan State University. *Journal of Educational Measurement*, 1971, *8*(4), 263-65.

Katz, I. Experimental studies of Negro-white relationships. In Berkowitz, L. (ed.), *Advances in experimental social psychology*, Vol. 5. New York: Academic Press, 1970.

Lesser, G.S. Cultural differences in learning and thinking styles. Harvard University, Laboratory of Human Development, 1972. Unpublished manuscript.

Lesser, G.S., Fifer, G., and Clark, D.H. Mental abilities of children from different social-class and cultural groups. *Monographs of the Society for Research in Child Development*, 1965, *30*(4).

McGraw-Hill Public Information and Publicity Department. Guidelines for equal treatment of the sexes in McGraw-Hill Book Company publications. New York: McGraw-Hill Book Company (undated).

McMorris, R.F., Brown, J.A., Snyder, G.W., and Pruzek, R.M. Effects of violating item construction principles. *Journal of Educational Measurement*, 1972, *9*, 287-95.

Mercer, J.R. *Labeling the mentally retarded*. Berkeley: University of California Press, 1973.

Messick, S., and Anderson, S. Educational testing, individual development, and social responsibility. *Counseling Psychologist*, 1970, *2*(2), 80-88.

National Association for the Advancement of Colored People. Resolutions of the 1974 Convention, New Orleans, Louisiana.

Newsweek. Quotas: The sleeper issue of '72? September 18, 1972, 36.

Pfeifer, C.M., Jr., and Sedlacek, W.E. The validity of academic predictors for black and white students at a predominantly white university. *Journal of Educational Measurement*, 1971, *8*(4), 253-61.

Schmidt, F.L., Berner, J.G., and Hunter, J.E. Racial differences in validity of employment tests: Reality or illusion? *Journal of Applied Psychology*, 1973, *58*, 5-9.

Smith, A.Z. (ed.). Use of NTE scores. *Education Recaps*, 1974, *14*(1), 3.

Singer, J.W. Employment report—U.S. readies new anti-bias guidelines. *National Journal Reports*, September 14, 1974.

Stodolsky, S.S., and Lesser, G.S. Learning patterns in the disadvantaged. *Harvard Education Review*, 1967, *37*(4), 546-93.

Thomas, A., Chess, S., and Birch, H. *Temperament and behavior disorders in children*. New York: New York University Press, 1968.

Thorndike, R.L. Concepts of culture-fairness. *Journal of Educational Measurement*, 1971, *8*(2), 63-70.

Time. Quarrel over quotas. October 9, 1972, 23-24.

Williams, R.L. The problem of match and mismatch in testing black children. In Miller, L.P. (ed.), *The testing of black students: A symposium*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1974.